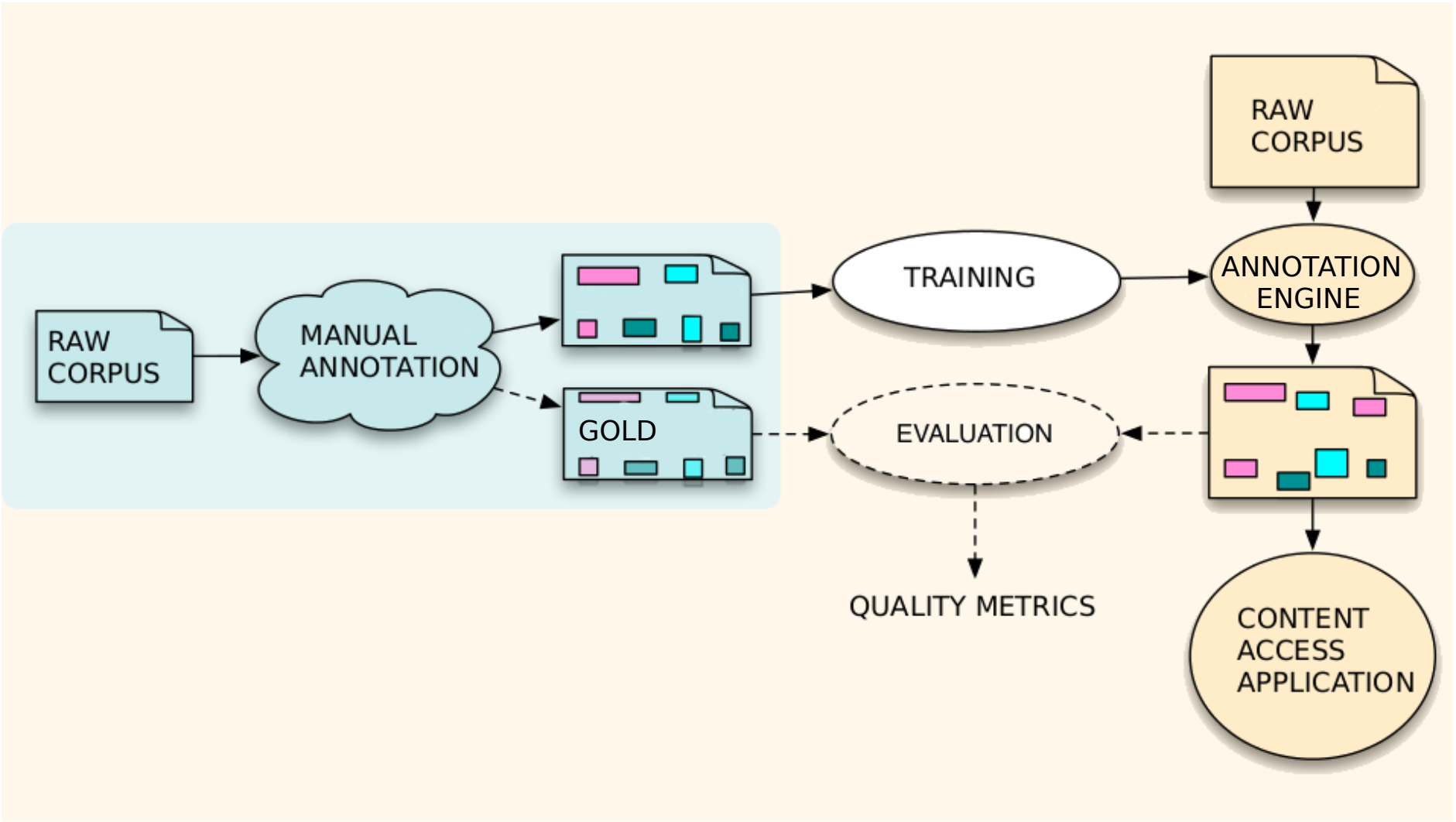
Crowdsourcing and human annotation: going beyond the legends to create quality data

Karën Fort

February 27th, 2015

Annotated corpora in Natural Language Processing (NLP)



Manual annotation: notoriously costly

Penn Treebank [Marcus et al., 1993]:

- 4.8 million tokens annotated with POS ⇒ learning phase of 1 month, to reach 3,000 words/h
- 3 million tokens annotated in syntax ⇒ learning phase of 2 months, to reach 475 words/h

Prague Dependency Treebank [Böhmová et al., 2001]:

- 1.8 million tokens annotated with POS and syntax

⇒ 5 years, 22 persons (max. 17 in parallel), $600,000

Manual annotation: notoriously costly

GENIA [Kim et al., 2008]:

- 9,372 sentences annotated in microbiology (proteins and gene names)

$\Rightarrow$ 5 part-time annotators, 1 senior coordinator and 1 junior for 1.5 year

CRAFT [Verspoor et al., 2012]:

- nearly 800,000 tokens annotated in POS, syntax and named entities in microbiology
- 3 years, approx. \$450,000 in annotation only

A view on crowdsourcing

`Wikipedia`, `Gutenberg Project`:

- benevolent (no remuneration)
- direct (the purpose is known)

`Games With A Purpose` (GWAPs):

- benevolent (no remuneration)
- indirect (the purpose is more or less hidden)

`Amazon Mechanical Turk` (AMT):

- remunerated
- direct

See [Geiger et al., 2011] for a detailed state of the art of the crowdsourcing taxonomies

A view on crowdsourcing

`Wikipedia`, `Gutenberg Project`:

- benevolent (no remuneration)
- direct (the purpose is known)

`Games With A Purpose` (GWAPs):

- benevolent (no remuneration)
- indirect (the purpose is more or less hidden)

`Amazon Mechanical Turk` (AMT):

- remunerated
- direct

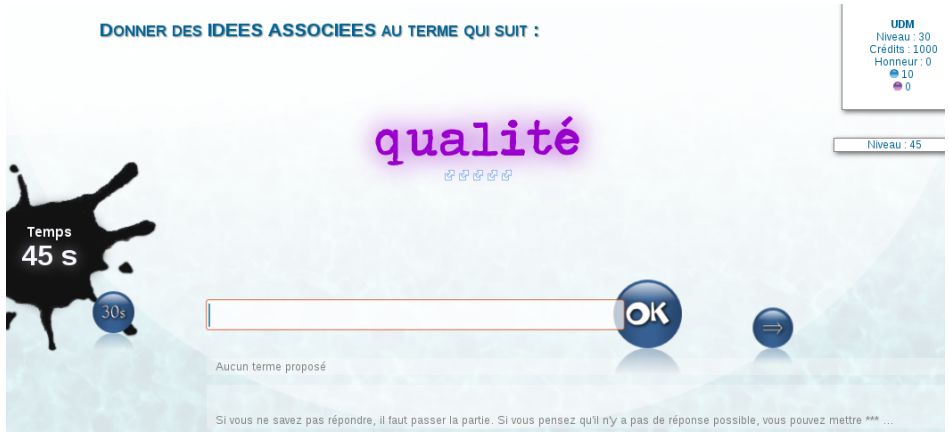
See [Geiger et al., 2011] for a detailed state of the art of the crowdsourcing taxonomies

JeuxDeMots: playing association of ideas...

...to create a lexical network [Lafourcade and Joubert, 2008]

More than 10 million relations created, that are constantly updated

- play by pairs
- more and more complex, typed relations
- challenges
- lawsuits, etc.

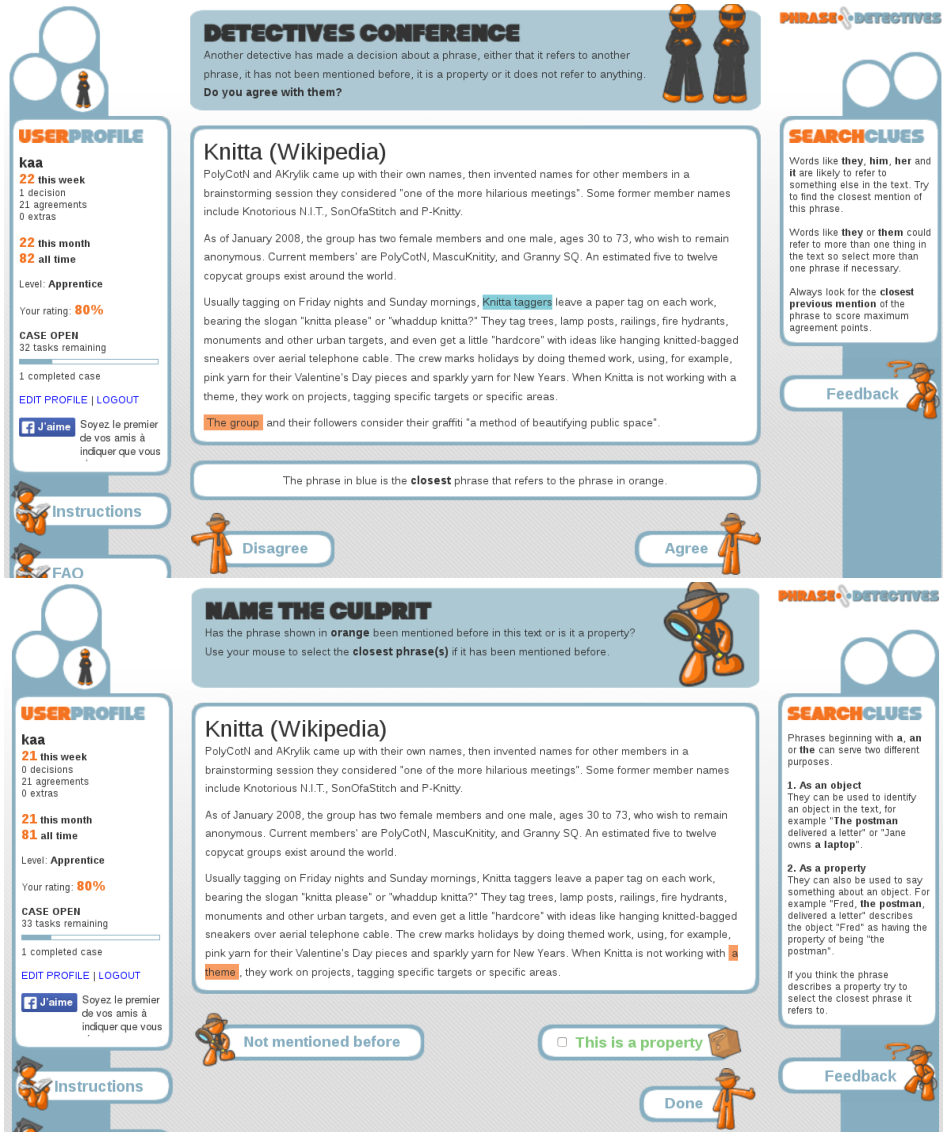


Phrase Detectives: playing detective...

...to annotate co-reference [Chamberlain et al., 2008]

200,000 words annotated corpus:

- pre-annotated
- detailed instructions
- training
- 2 different playing modes:
  - annotation
  - validation (correction of annotations)



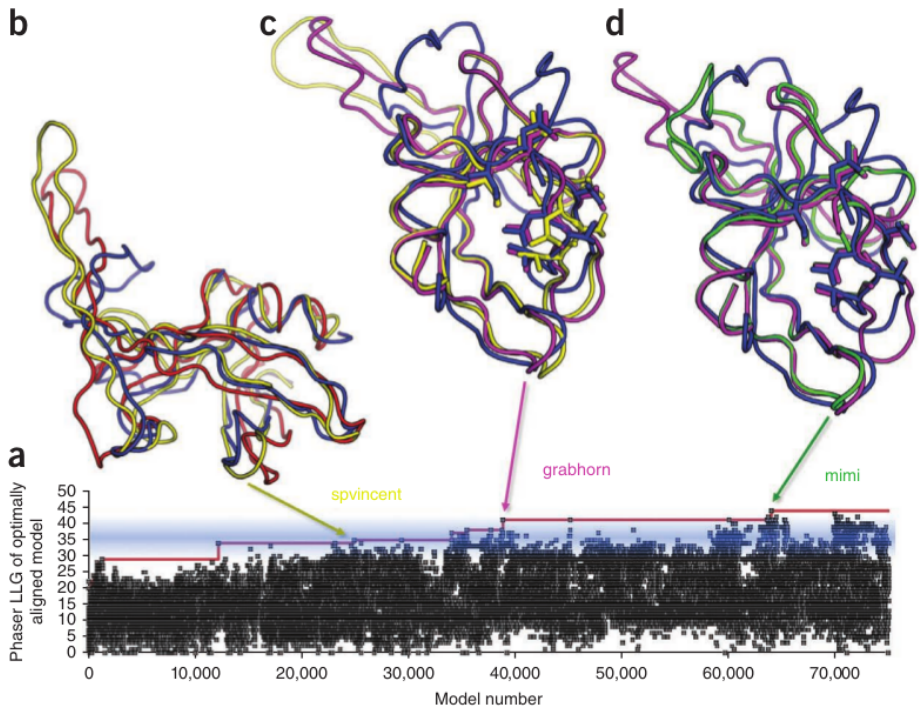
FoldIt: playing proteins folding...

...to solve scientific issues [Khatib et al., 2011]

Solution to the crystal structure of a monomeric retroviral protease (simian AIDS-causing monkey virus)

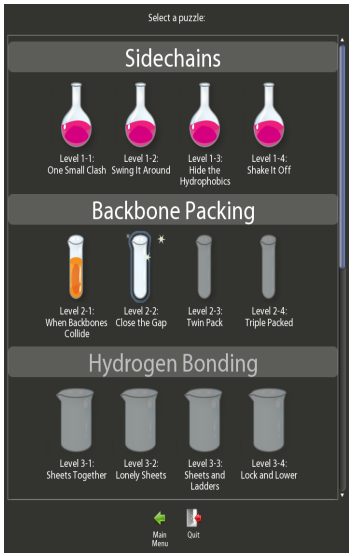
Solution to an issue unsolved for over a decade:

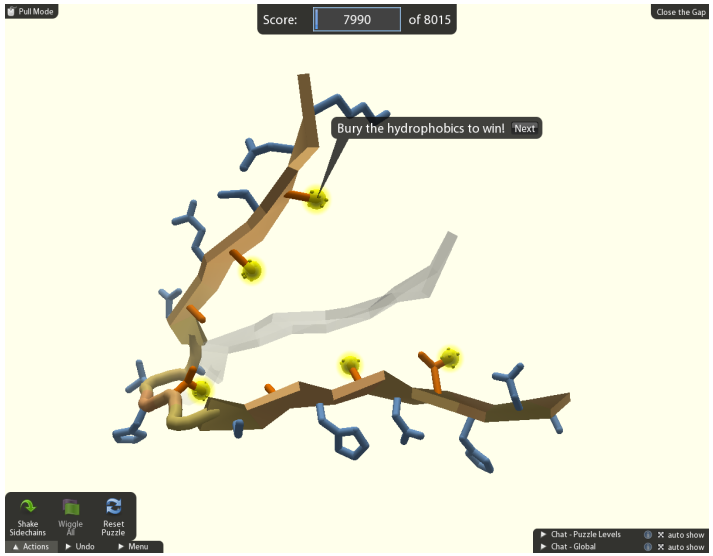
- found in a couple of weeks
- step by step
- by a team of players
- that will allow for the creation of antiretroviral drugs



FoldIt: playing proteins folding...

...without any prior knowledge in biochemistry [Cooper et al., 2010]



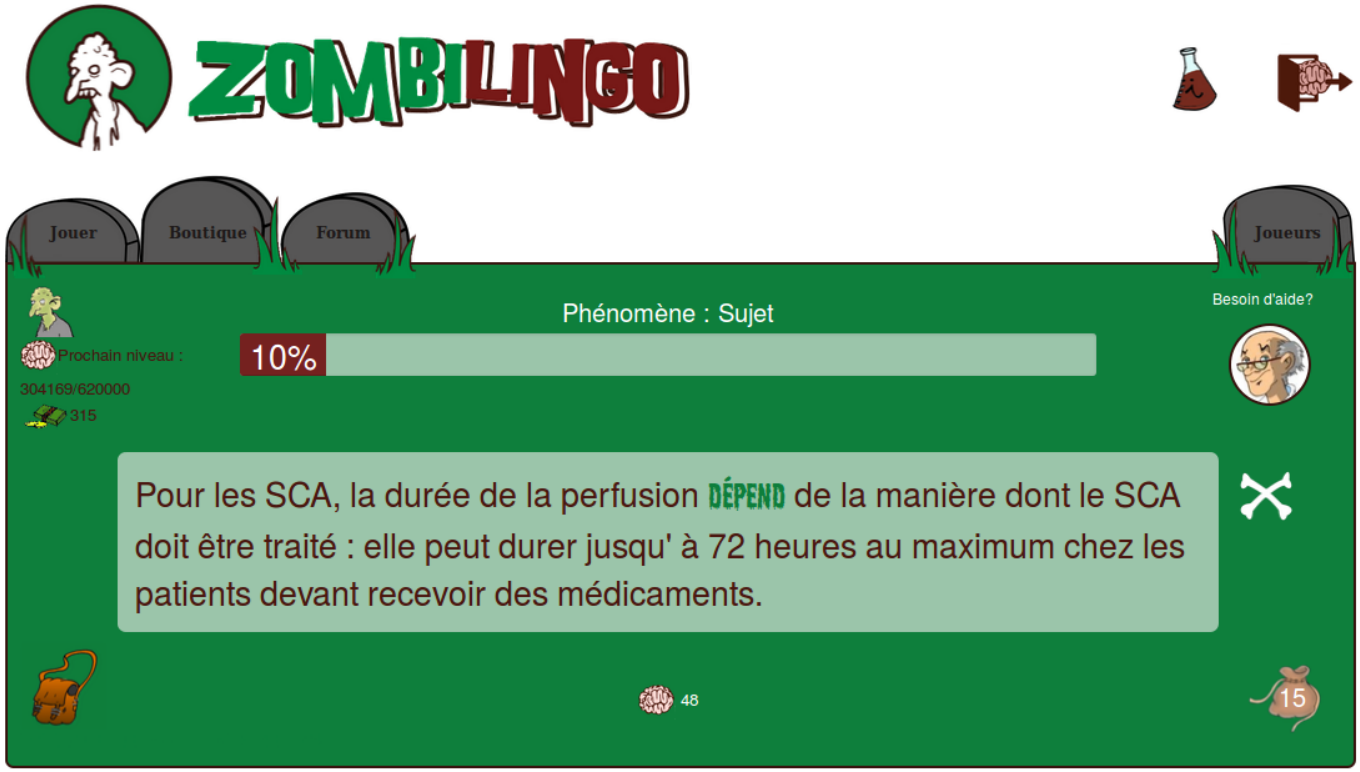


Step-by-step training:

- tutorial decomposed by concepts
- puzzles of each concept
- access to the following puzzles is given only if your level is sufficient

ZombiLingo: eating heads...

...to annotate (French) corpora with dependency syntax [Fort et al., 2014]



V 1.0 being finalized...

- decomposition of the task by phenomenon (not by sentence)
- tutorial by phenomenon
- regularly proposed reference sentences

Going beyond legends

A promising solution:

- players like (love?) to follow rules!
- massive and quick
- (relatively) low cost
- various productions (limits?)
- creation of dynamic language resources

Going beyond legends

A promising solution:

- players like (love?) to follow rules!
- massive and quick
- (relatively) low cost
- various productions (limits?)
- creation of dynamic language resources

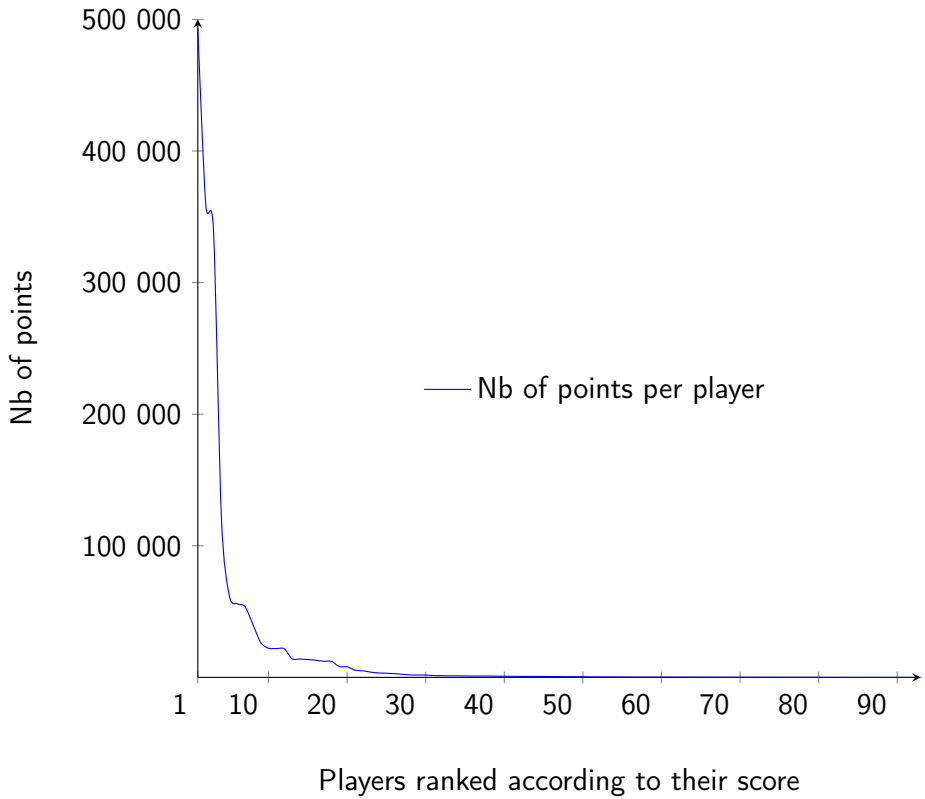
(Still) little studied, need to:

→ deconstruct the legends (myths?)

→ evaluate the quality of the produced resources

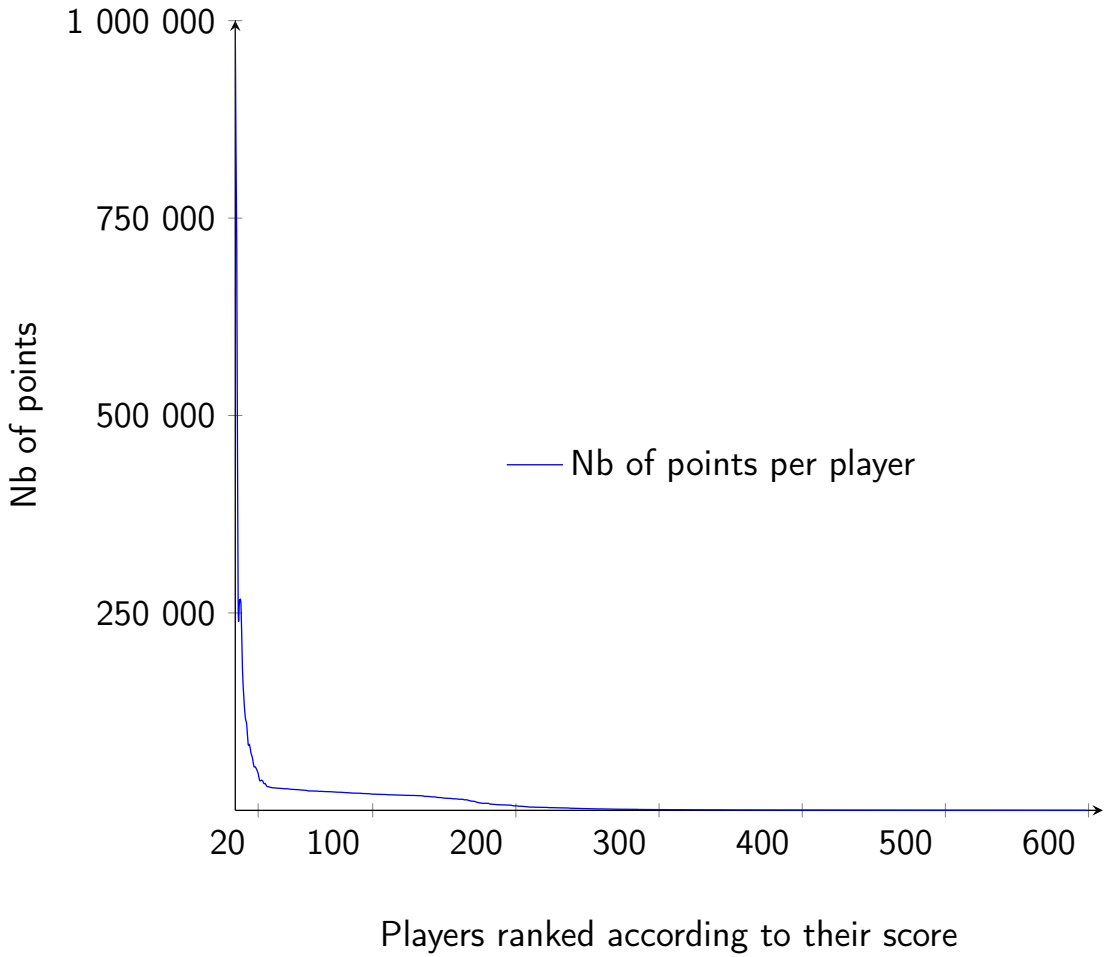
→ identify the complexity of an annotation task to be able to reduce it

A crowd of "non-experts"? (GWAP)



Players on `Phrase Detectives` (Feb. 2011 - Feb. 2012) [Chamberlain et al., 2013]

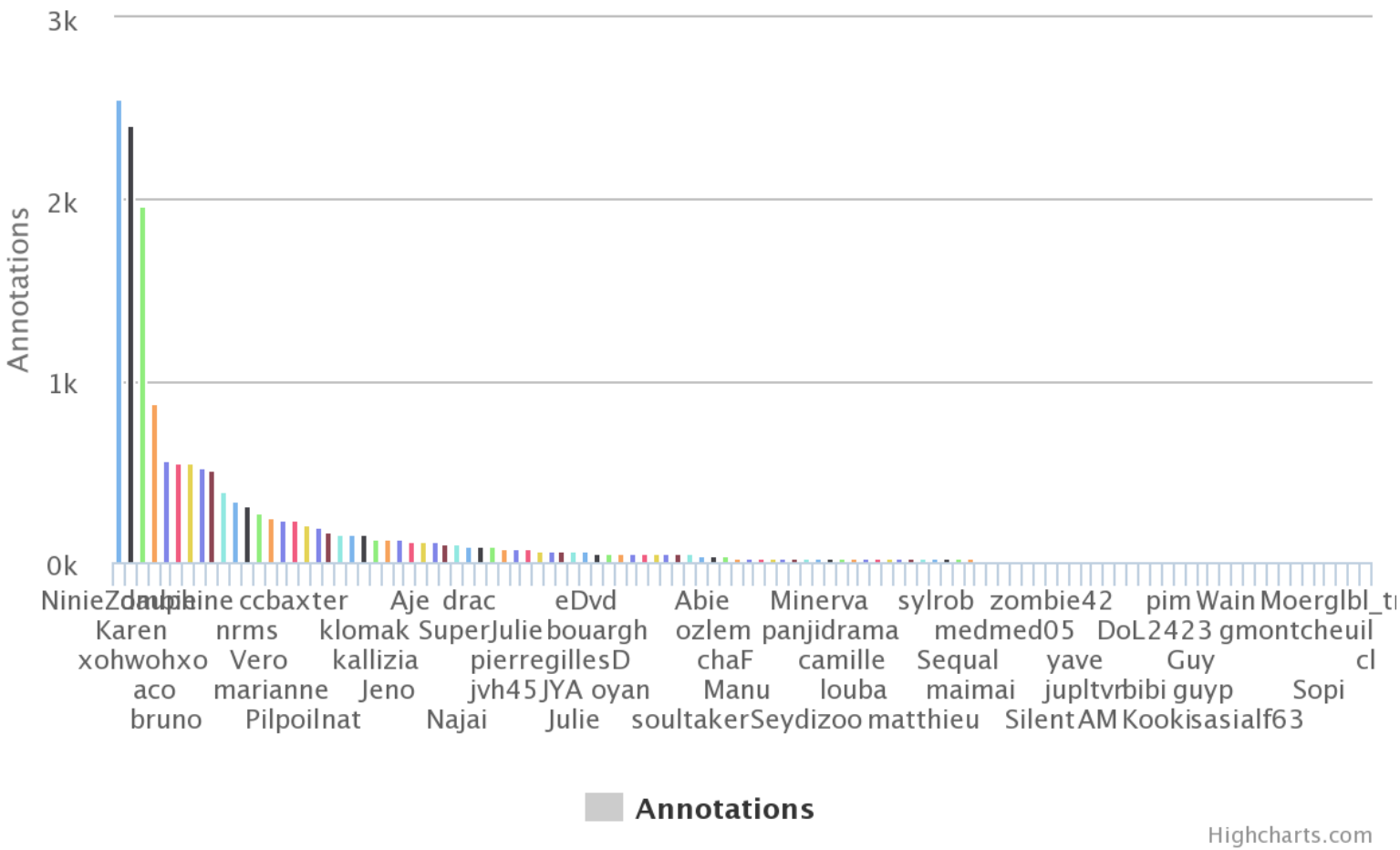
A crowd of "non-experts"? (GWAP (2))



Players on JeuxDeMots

(source : http://www.jeuxdemots.org/generateRanking-4.php)

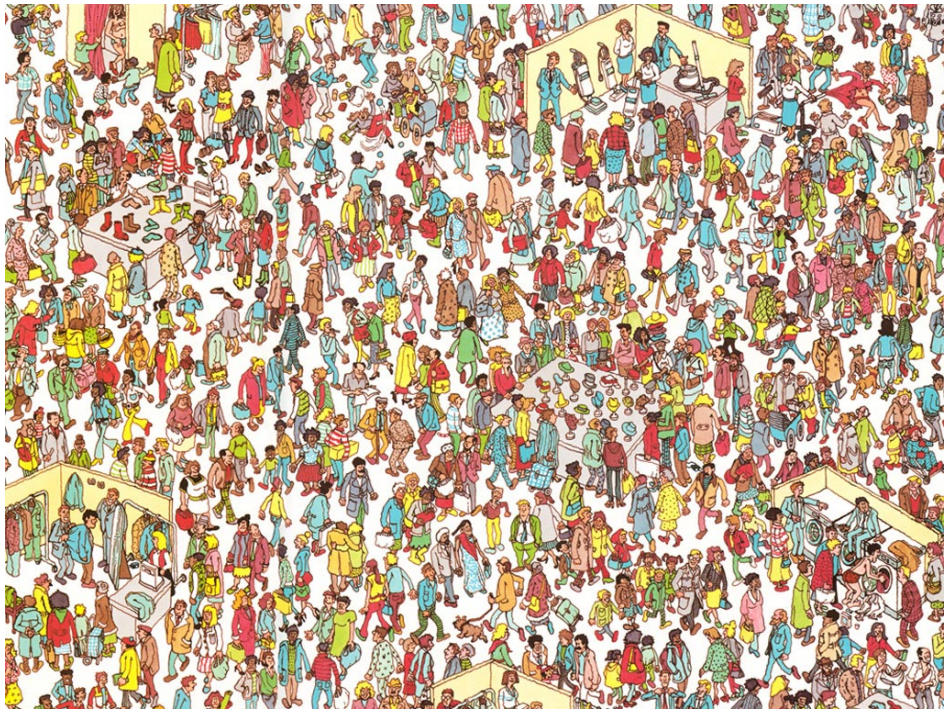
A crowd of "non-experts"? (GWAP (3))



Nb of annotations per player on ZombiLingo (Feb. 2015)

Crowdsourcing annotation

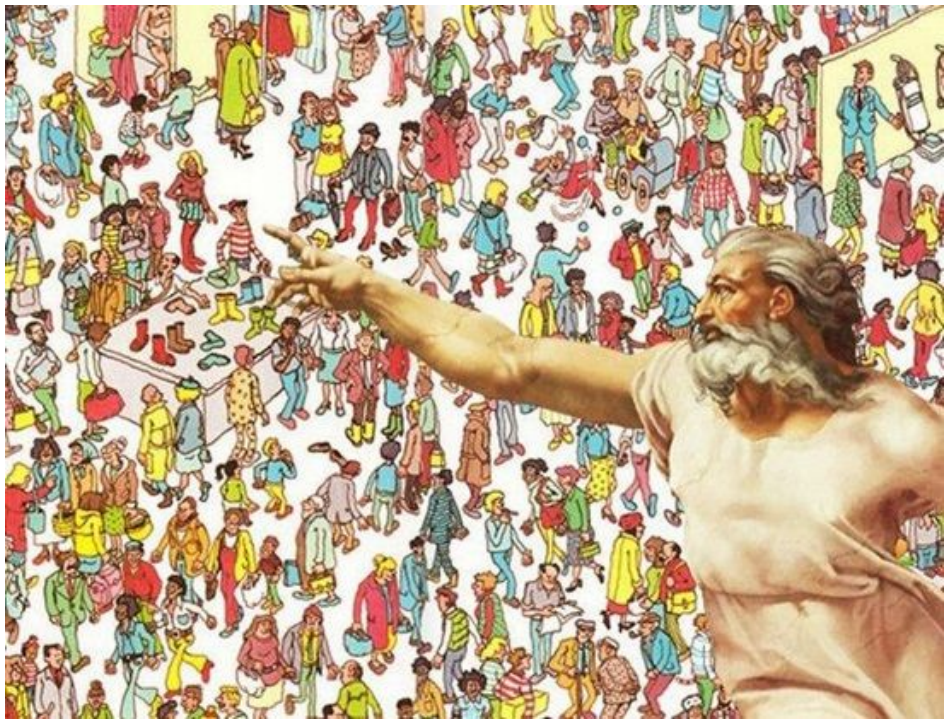
Production of annotations by "non-experts"?



Crowdsourcing annotation

~~Production of annotations by "non-experts"?~~

→ Find/train experts (of the task) in the crowd



Creating quality data *vs* creating game features

preserving the virtuous circle is not always straightforward

Creating quality data *vs* creating game features

preserving the virtuous circle is not always straightforward

sentence that disappears in ZombiLingo:

+ the player is surprised: fun!
- the player clicks anywhere: creation of a bad quality resource

Creating quality data *vs* creating game features

preserving the virtuous circle is not always straightforward

sentence that disappears in ZombiLingo:

+ the player is surprised: fun!
- the player clicks anywhere: creation of a bad quality resource

player who found a hack in JeuxDeMots's code to get more time:

+ creates more good data: creation of a good quality resource
- generates envy and anger in the community of players: bad for the game

Quality of the created resource? Phrase Detectives

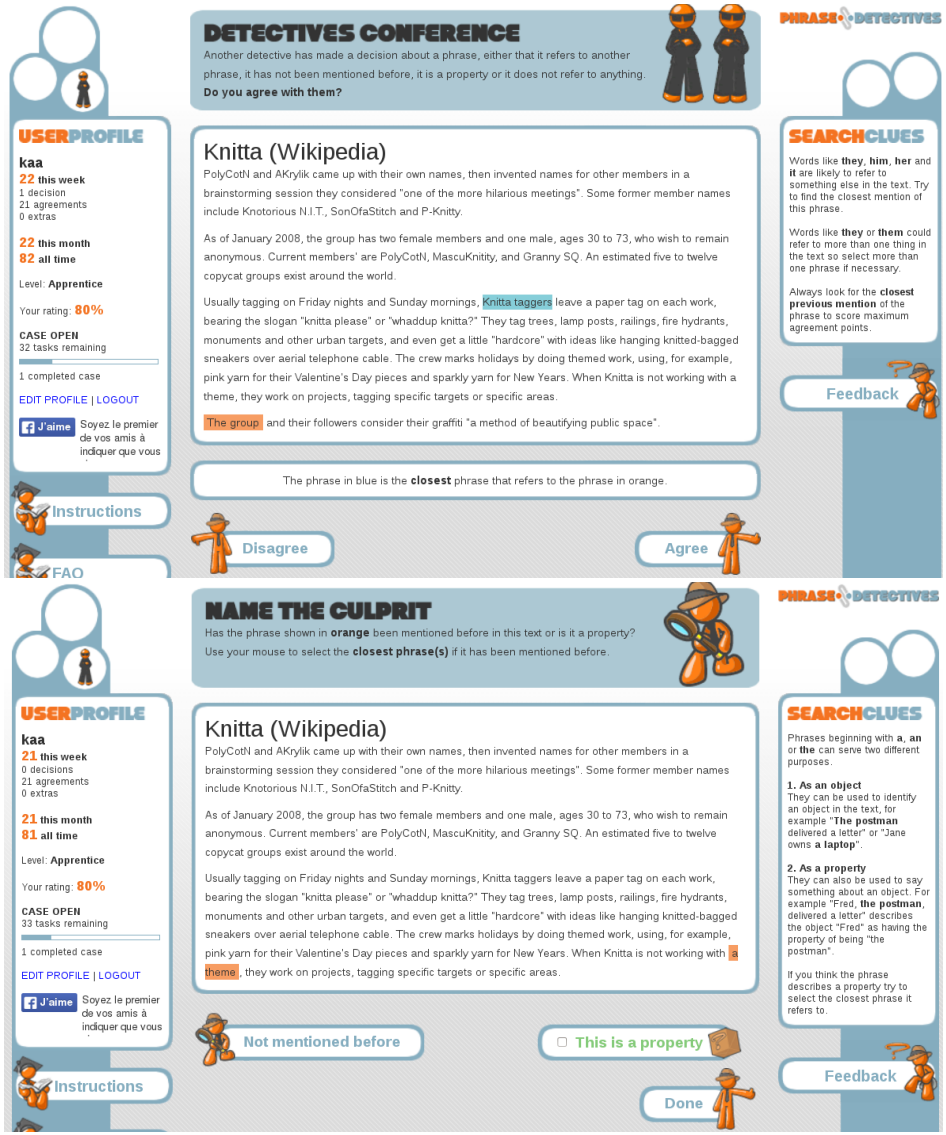
using a reference

Evaluation:

- reference corpus
- high observed inter-annotator agreement (from 0.7 to 0.8) [Chamberlain et al., 2009]

Failure: identification of properties

*Jon, the **postman**, delivered the letter*

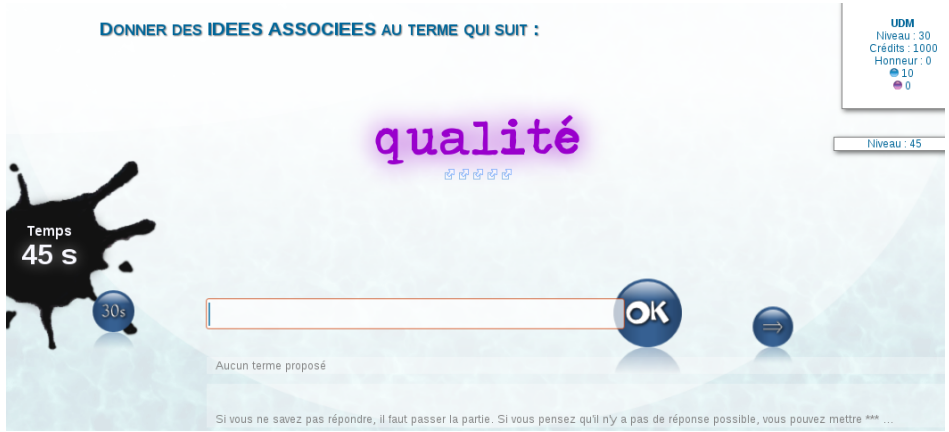


Quality of the created resource? JeuxDeMots

using another game!

- no (real) reference (though Babelnet...)

⇒ creation of a game to validate the resource
[Lafourcade et al., 2011]



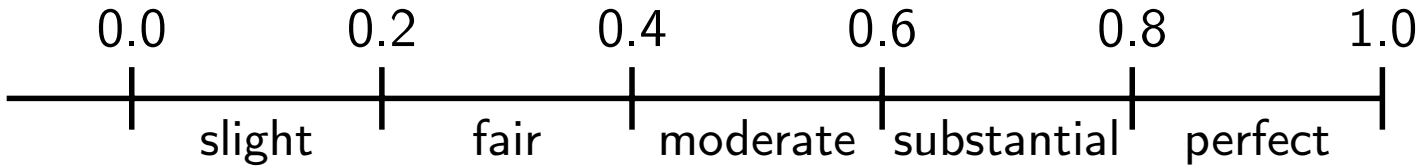
Evaluating human interpretation?

We can only measure the **consistency** of annotation i.e.
if humans make **consistent** decisions
taking **chance** into account

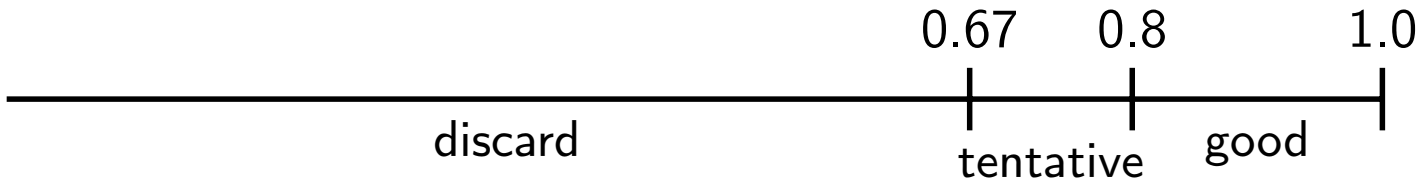
$$\kappa$$

Scales for the interpretation of Kappa

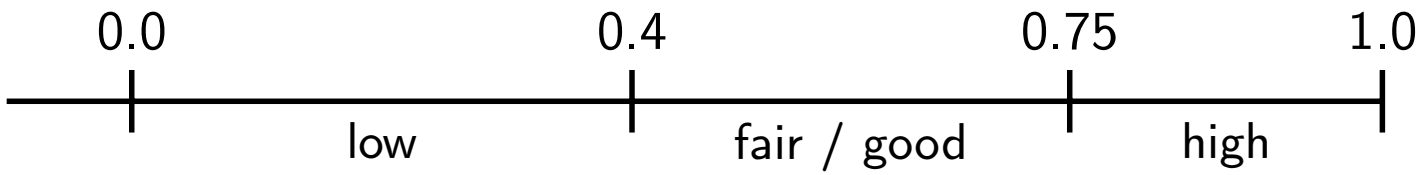
- Landis and Koch, 1977



- Krippendorff, 1980



- Green, 1997



- "if a threshold needs to be set, 0.8 is a good value" [Artstein and Poesio, 2008]

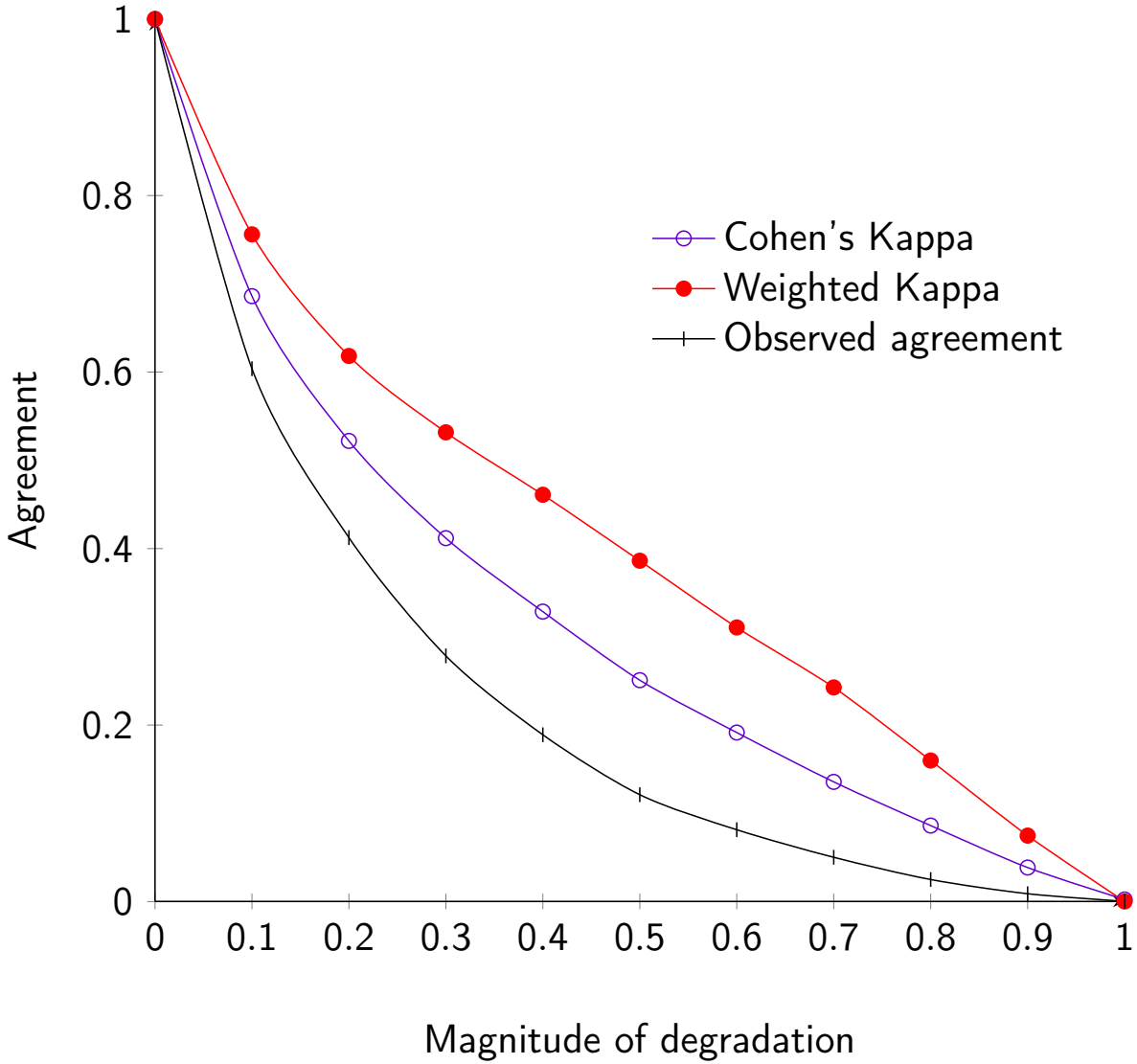
Giving meaning to the obtained results [Mathet et al., 2012]

`Richter` tool that:

- input: a reference annotation (real or generated automatically)
- generates degradations of a certain magnitude (from 0 to 1)
- applies one or several inter-annotator agreement measures on each set of annotations (corresponding to a magnitude of degradation)

Richter on the TCOF-POS corpus [Benzitoun et al., 2012]

Prevalence not taken into account, but proximity between categories is:



What is difficult? How to help?

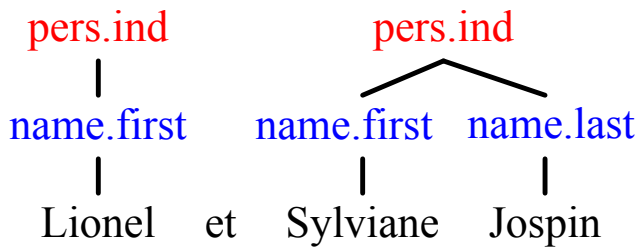
Part-of-speech [Marcus et al., 1993] :

I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.

Gene renaming relations [Fort et al., 2012a] :

The yppB gene complemented the defect of the recG40 strain. yppB and ypbC and their respective null alleles were termed recU and "recU1" (recU:cat) and recS and "recS1" (recS:cat), respectively. The recU and recS mutations were introduced into rec-deficient strains representative of the alpha (recF), beta (addA5 addB72), gamma (recH342), and epsilon (recG40) epistatic groups.

Structured named entities [Grouin et al., 2011] :



A growing interest in the community

- Large-scale campaigns feedback
  [Marcus et al., 1993, Abeillé et al., 2003]

A growing interest in the community

- Large-scale campaigns feedback
  [Marcus et al., 1993, Abeillé et al., 2003]
- Good practices:

A growing interest in the community

- Large-scale campaigns feedback
  [Marcus et al., 1993, Abeillé et al., 2003]
- Good practices:
  - formats [Ide and Romary, 2006]

A growing interest in the community

- Large-scale campaigns feedback
  [Marcus et al., 1993, Abeillé et al., 2003]
- Good practices:
  - formats [Ide and Romary, 2006]
  - organization [Bontcheva et al., 2010]

A growing interest in the community

- Large-scale campaigns feedback
  [Marcus et al., 1993, Abeillé et al., 2003]
- Good practices:
  - formats [Ide and Romary, 2006]
  - organization [Bontcheva et al., 2010]
  - evaluation [Krippendorff, 2004]

A growing interest in the community

- Large-scale campaigns feedback
  [Marcus et al., 1993, Abeillé et al., 2003]
- Good practices:
  - formats [Ide and Romary, 2006]
  - organization [Bontcheva et al., 2010]
  - evaluation [Krippendorff, 2004]
- Partial methodologies:

A growing interest in the community

- Large-scale campaigns feedback
  [Marcus et al., 1993, Abeillé et al., 2003]
- Good practices:
  - formats [Ide and Romary, 2006]
  - organization [Bontcheva et al., 2010]
  - evaluation [Krippendorff, 2004]
- Partial methodologies:
  - tutorial by E. Hovy (ACL 2010),

A growing interest in the community

- Large-scale campaigns feedback
  [Marcus et al., 1993, Abeillé et al., 2003]
- Good practices:
  - formats [Ide and Romary, 2006]
  - organization [Bontcheva et al., 2010]
  - evaluation [Krippendorff, 2004]
- Partial methodologies:
  - tutorial by E. Hovy (ACL 2010),
  - agile annotation
    [Bonneau-Maynard et al., 2005, Voormann and Gut, 2008],

A growing interest in the community

- Large-scale campaigns feedback
  [Marcus et al., 1993, Abeillé et al., 2003]
- Good practices:
  - formats [Ide and Romary, 2006]
  - organization [Bontcheva et al., 2010]
  - evaluation [Krippendorff, 2004]
- Partial methodologies:
  - tutorial by E. Hovy (ACL 2010),
  - agile annotation
    [Bonneau-Maynard et al., 2005, Voormann and Gut, 2008],
  - MATTER [Pustejovsky and Stubbs, 2012], light annotation
    [Stubbs, 2012]

A growing interest in the community

- Large-scale campaigns feedback
  [Marcus et al., 1993, Abeillé et al., 2003]
- Good practices:
  - formats [Ide and Romary, 2006]
  - organization [Bontcheva et al., 2010]
  - evaluation [Krippendorff, 2004]
- Partial methodologies:
  - tutorial by E. Hovy (ACL 2010),
  - agile annotation
    [Bonneau-Maynard et al., 2005, Voormann and Gut, 2008],
  - MATTER [Pustejovsky and Stubbs, 2012], light annotation
    [Stubbs, 2012]
- Some insights from cognitive science [Tomanek et al., 2010]

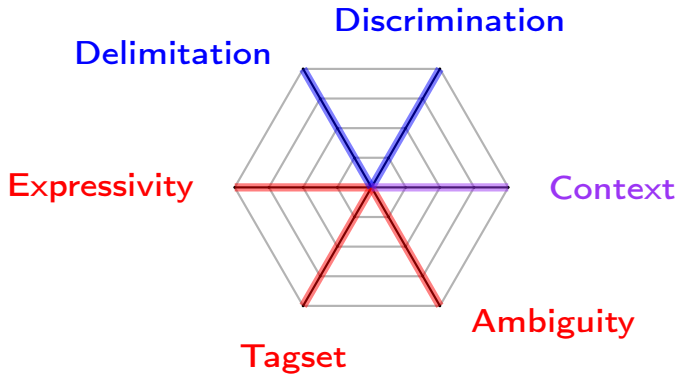
A growing interest in the community

- Large-scale campaigns feedback
  [Marcus et al., 1993, Abeillé et al., 2003]
- Good practices:
  - formats [Ide and Romary, 2006]
  - organization [Bontcheva et al., 2010]
  - evaluation [Krippendorff, 2004]
- Partial methodologies:
  - tutorial by E. Hovy (ACL 2010),
  - agile annotation
    [Bonneau-Maynard et al., 2005, Voormann and Gut, 2008],
  - MATTER [Pustejovsky and Stubbs, 2012], light annotation
    [Stubbs, 2012]
- Some insights from cognitive science [Tomanek et al., 2010]

What is complex in manual annotation?

Complexity dimensions [Fort et al., 2012b]

- 5 independent dimensions:
  - 2 related to the localisation of annotations
  - 3 related to the characterisation of annotations
- 1 not independent: the context



- Scale from **0** (null complexity) to **1** (maximal complexity) to allow for the comparison between campaigns
- Independent from the volume to annotate and the number of annotators

Discrimination

Parts-of-speech [Marcus et al., 1993], pre-annotated :

I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.

Gene renaming[Fort et al., 2012a], no pre-annotation:

The yppB:cat and ypbC:cat null alleles rendered cells sensitive to DNA-damaging agents, impaired plasmid transformation (25- and 100-fold), and moderately affected chromosomal transformation when present in an otherwise Rec+ B. subtilis strain. The yppB gene complemented the defect of the recG40 strain. yppB and ypbC and their respective null alleles were termed recU and "recU1" (recU:cat) and recS and "recS1" (recS:cat), respectively. The recU and recS mutations were introduced into rec-deficient strains representative of the alpha (recF), beta (addA5 addB72), gamma (recH342), and epsilon (recG40) epistatic groups.

Discrimination

Parts-of-speech [Marcus et al., 1993], pre-annotated :

I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.

Gene renaming[Fort et al., 2012a], no pre-annotation:

The yppB:cat and ypbC:cat null alleles rendered cells sensitive to DNA-damaging agents, impaired plasmid transformation (25- and 100-fold), and moderately affected chromosomal transformation when present in an otherwise Rec+ B. subtilis strain. The yppB gene complemented the defect of the recG40 strain. yppB and ypbC and their respective null alleles were termed recU and "recU1" (recU:cat) and recS and "recS1" (recS:cat), respectively. The recU and recS mutations were introduced into rec-deficient strains representative of the alpha (recF), beta (addA5 addB72), gamma (recH342), and epsilon (recG40) epistatic groups.

$\Rightarrow$ **more difficult** if the units to annotate are scattered, in particular if the segmentation is not obvious.

Discrimination

The discrimination weight is all the more high as the proportion of what *should* be annotated as compared to what *could* be annotated is low.

Definition

$$Discrimination(Flow) = 1 - \frac{|Annotations(Flow)|}{\sum_{i=1}^{LevelSeg} |UnitsObtainedBySeg_i(Flow)|}$$

$\Rightarrow$ Need for a reference segmentation

Parts-of-speech[Marcus et al., 1993] :

I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.

$$Discrimination_{PTB_{POS}} = 0$$

Gene renaming[Fort et al., 2012a] :

The yppB:cat and ypbC:cat null alleles rendered cells sensitive to DNA-damaging agents, impaired plasmid transformation (25- and 100-fold), and moderately affected chromosomal transformation when present in an otherwise Rec+ B. subtilis strain. The yppB gene complemented the defect of the recG40 strain. yppB and ypbC and their respective null alleles were termed recU and "recU1" (recU:cat) and recS and "recS1" (recS:cat), respectively. The recU and recS mutations were introduced into rec-deficient strains representative of the alpha (recF), beta (addA5 addB72), gamma (recH342), and epsilon (recG40) epistatic groups.

$$Discrimination_{Renaming} = 0,95$$

Boundaries delimitation

- **extending** or **shrinking** the discriminated unit:
  Madame Chirac $\rightarrow$ *Monsieur et Madame Chirac*

Boundaries delimitation

- **extending** or **shrinking** the discriminated unit:
  Madame Chirac → *Monsieur et Madame Chirac*
- **decompose** a discriminated unit into several elements:
  le préfet Érignac → *le préfet Érignac*

Boundaries delimitation

- **extending** or **shrinking** the discriminated unit:
  Madame Chirac → *Monsieur et Madame Chirac*
- **decompose** a discriminated unit into several elements:
  le préfet Érignac → *le préfet Érignac*
- or **group** together several discriminated units into one unique annotation:
  Sa Majesté
  le roi Mohamed VI → *Sa Majesté le roi Mohamed VI*

Boundaries delimitation

Definition

$$Delimitation(Flow) = min\left(\frac{Substitutions + Additions + Deletions}{|Annotations(Flow)|}, 1\right)$$

$$Delimitation_{Renaming} = 0$$

$$Delimitation_{PTB_{POS}} = 0$$

$$Délimitation_{EN_{TypesSubtypes}} = 1$$

Expressiveness of the annotation language

Definition

The degrees of expressiveness of the annotation language are the following:

- 0.25: type languages
- 0.5: relational languages of arity 2
- 0.75: relational languages of arity higher than 2
- 1: higher-order languages

$$Expressiveness_{Renaming} = 0.25$$

Dimension of the tagset

<table>
<tr><th colspan="3">Person</th><th colspan="3">Function</th></tr>
<tr><td colspan="2">pers.ind (individual person)</td><td>pers.coll (group of persons)</td><td colspan="2">func.ind (individual function)</td><td>func.coll (collectivity of functions)</td></tr>
<tr><th colspan="3">Location</th><th colspan="3">Production</th></tr>
<tr><td>administrative (loc.adm.town, loc.adm.reg, loc.adm.nat, loc.adm.sup)</td><td>physical (loc.phys.geo, loc.phys.hydro, loc.phys.astro)</td><td>facilities (loc.fac), oronyms (loc.oro), address (loc.add.phys, loc.add.elec)</td><td>prod.object (manufactured object)</td><td>prod.serv (transportation route)</td><td>prod.fin (financial products)</td></tr>
<tr><td colspan="3"></td><td>prod.doctr (doctrine)</td><td>prod.rule (law)</td><td>prod.soft (software)</td></tr>
<tr><td colspan="3"></td><td>prod.art</td><td>prod.media</td><td>prod.award</td></tr>
<tr><th colspan="3">Organization</th><th colspan="3" rowspan="3">Time</th></tr>
<tr><td colspan="2">org.adm (administration)</td><td>org.ent (services)</td></tr>
<tr><th colspan="3">Amount</th></tr>
<tr><td colspan="3">amount (with unit or general object), including duration</td><td colspan="2">time.date.abs (absolute date), time.date.rel (relative date)</td><td>time.hour.abs (absolute hour), time.hour.rel (relative hour)</td></tr>
</table>

Types and sub-types used for structured NE annotation [Grouin et al., 2011]

Dimension of the tagset

<table>
<tr><td colspan="3">Person</td><td colspan="3">Function</td></tr>
<tr><td colspan="2">pers.ind (individual person)</td><td>pers.coll (group of persons)</td><td colspan="2">func.ind (individual function)</td><td>func.coll (collectivity of functions)</td></tr>
<tr><td colspan="3">Location</td><td colspan="3">Production</td></tr>
<tr><td>administrative (loc.adm.town, loc.adm.reg, loc.adm.nat, loc.adm.sup)</td><td>physical (loc.phys.geo, loc.phys.hydro, loc.phys.astro)</td><td>facilities (loc.fac), oronyms (loc.oro), address (loc.add.phys, loc.add.elec)</td><td>prod.object (manufactured object)</td><td>prod.serv (transportation route)</td><td>prod.fin (financial products)</td></tr>
<tr><td></td><td></td><td></td><td>prod.doctr (doctrine)</td><td>prod.rule (law)</td><td>prod.soft (software)</td></tr>
<tr><td></td><td></td><td></td><td>prod.art</td><td>prod.media</td><td>prod.award</td></tr>
<tr><td colspan="3">Organization</td><td colspan="3" rowspan="3">Time</td></tr>
<tr><td colspan="2">org.adm (administration)</td><td>org.ent (services)</td></tr>
<tr><td colspan="3">Amount</td></tr>
<tr><td colspan="3">amount (with unit or general object), including duration</td><td colspan="2">time.date.abs (absolute date), time.date.rel (relative date)</td><td>time.hour.abs (absolute hour), time.hour.rel (relative hour)</td></tr>
</table>

Level 1: *pers*, *func*, *loc*, *prod*, *org*, *time*, *amount* $\rightarrow$ 7 possibilities (degree of freedom = 6).

Dimension of the tagset

<table>
<tr><th colspan="3">Person</th><th colspan="3">Function</th></tr>
<tr><td colspan="2">pers.ind (individual person)</td><td>pers.coll (group of persons)</td><td colspan="2">func.ind (individual function)</td><td>func.coll (collectivity of functions)</td></tr>
<tr><th colspan="3">Location</th><th colspan="3">Production</th></tr>
<tr><td>administrative (loc.adm.town, loc.adm.reg, loc.adm.nat, loc.adm.sup)</td><td>physical (loc.phys.geo, loc.phys.hydro, loc.phys.astro)</td><td>facilities (loc.fac), oronyms (loc.oro), address (loc.add.phys, loc.add.elec)</td><td>prod.object (manufac-tured object)</td><td>prod.serv (transporta-tion route)</td><td>prod.fin (financial products)</td></tr>
<tr><td colspan="3"></td><td>prod.doctr (doctrine)</td><td>prod.rule (law)</td><td>prod.soft (software)</td></tr>
<tr><td colspan="3"></td><td>prod.art</td><td>prod.media</td><td>prod.award</td></tr>
<tr><th colspan="3">Organization</th><th colspan="3">Time</th></tr>
<tr><td colspan="2">org.adm (administra-tion)</td><td>org.ent (services)</td><td colspan="2" rowspan="2">time.date.abs (absolute date), time.date.rel (relative date)</td><td rowspan="2">time.hour.abs (absolute hour), time.hour.rel (relative hour)</td></tr>
<tr><th colspan="3">Amount</th></tr>
<tr><td colspan="3">amount (with unit or general object), includ-ing duration</td></tr>
</table>

Level 1: *pers*, *func*, *loc*, *prod*, *org*, *time*, *amount* $\rightarrow$ 7 possibilities (degree of freedom = 6).
Level 2: *prod.object*, *prod.serv*, *prod.fin*, *prod.soft*, *prod.doctr*, *prod.rule*, *prod.art*, *prod.media*, *prod.award* $\rightarrow$ 9 possibilities (degree of freedom = 8).

Dimension of the tagset

<table>
<tr><th colspan="3">Person</th><th colspan="3">Function</th></tr>
<tr><td colspan="2">pers.ind (individual person)</td><td>pers.coll (group of persons)</td><td colspan="2">func.ind (individual function)</td><td>func.coll (collectivity of functions)</td></tr>
<tr><th colspan="3">Location</th><th colspan="3">Production</th></tr>
<tr><td>administrative (loc.adm.town, loc.adm.reg, loc.adm.nat, loc.adm.sup)</td><td>physical (loc.phys.geo, loc.phys.hydro, loc.phys.astro)</td><td>facilities (loc.fac), oronyms (loc.oro), address (loc.add.phys, loc.add.elec)</td><td>prod.object (manufac-tured object)</td><td>prod.serv (transporta-tion route)</td><td>prod.fin (financial products)</td></tr>
<tr><td colspan="3"></td><td>prod.doctr (doctrine)</td><td>prod.rule (law)</td><td>prod.soft (software)</td></tr>
<tr><td colspan="3"></td><td>prod.art</td><td>prod.media</td><td>prod.award</td></tr>
<tr><th colspan="3">Organization</th><th colspan="3" rowspan="3">Time</th></tr>
<tr><td colspan="2">org.adm (administra-tion)</td><td>org.ent (services)</td></tr>
<tr><th colspan="3">Amount</th></tr>
<tr><td colspan="3">amount (with unit or general object), includ-ing duration</td><td colspan="2">time.date.abs (absolute date), time.date.rel (relative date)</td><td>time.hour.abs (absolute hour), time.hour.rel (relative hour)</td></tr>
</table>

Level 1: *pers*, *func*, *loc*, *prod*, *org*, *time*, *amount* $\rightarrow$ 7 possibilities (degree of freedom = 6).
Level 2: *prod.object*, *prod.serv*, *prod.fin*, *prod.soft*, *prod.doctr*, *prod.rule*, *prod.art*, *prod.media*, *prod.award* $\rightarrow$ 9 possibilities (degree of freedom = 8).
Level 3: *loc.adm.town*, *loc.adm.reg*, *loc.adm.nat*, *loc.adm.sup* $\rightarrow$ 4 possibilities (degree of freedom = 3).

Dimension of the tagset

Degree of freedom

$$\nu = \nu_1 + \nu_2 + \ldots + \nu_m$$

where $\underline{\nu_i}$ is the maximal degree of freedom the annotator has when choosing the $\underline{i^{th}}$ sub-type ($\underline{\nu_i = n_i - 1}$).

Dimension of the tagset

$$Dimension(Flow) = min(\frac{\nu}{\tau}, 1)$$

where τ is the threshold from which we consider the tagset to be very large (experimentally determined).

$$Dimension_{Renaming} = 0.04$$
$$Dimension_{NE_{TypesSubtypes}} = 0.34$$

Degree of ambiguity: residual ambiguity

Using the traces left by the annotators:

[...] <EukVirus>3CDproM</EukVirus> can process both structural and nonstructural precursors of the <EukVirus ***uncertainty-type*** *= "too-generic"><taxon>poliovirus</taxon> polyprotein</EukVirus> [...].*

Définition

$$AmbiguityRes(Flow) = \frac{|Annotations_{amb}|}{|Annotations|}$$

$$AmbiguityRes_{Renaming} = 0.02$$

$\rightarrow$ does not apply to the `Penn Treebank` (no traces).

Degree of ambiguity: theoretical ambiguity

Proportion of the units to annotate that corresponds to ambiguous vocables.

Definition

$$AmbiguityTh(Flow) = \frac{\sum_{voc_i=1}^{|Voc(Flow)|} (Ambig(voc_i) * freq(voc_i, Flow))}{|Units(Flow)|}$$

with

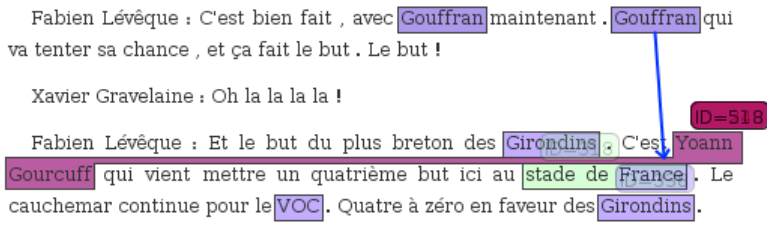
$$Ambig(voc_i) = \begin{cases} 1 & \text{if} \quad |Tags(voc_i)| > 1 \\ 0 & \text{else} \end{cases}$$

$\rightarrow$ Does not apply to renaming relations.

$$AmbiguityTh_{Identification} = 0.01$$

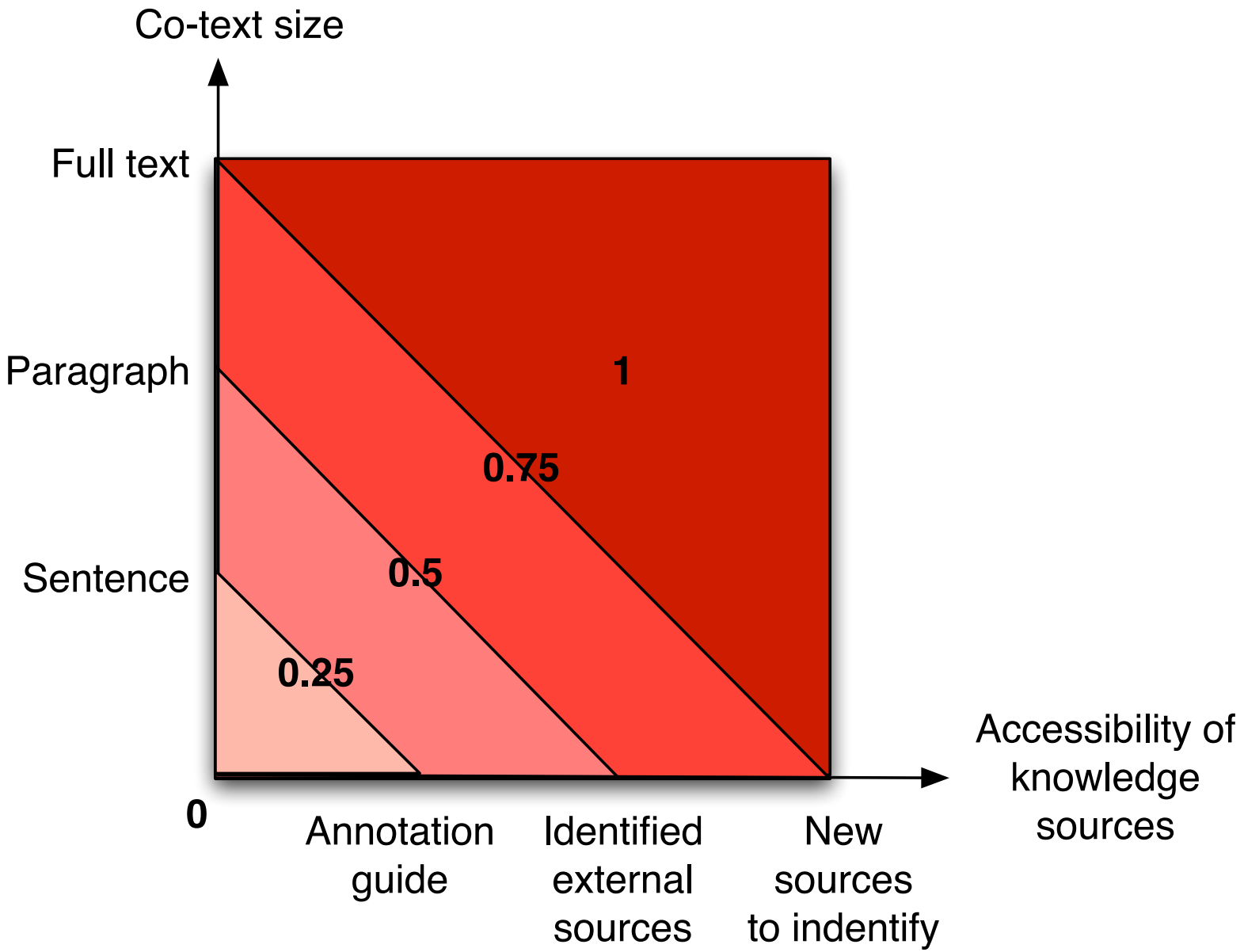
Context to take into account

- **size of the window** to take into account in the source signal:
  - The sentence:
    I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.
  - ... or more:



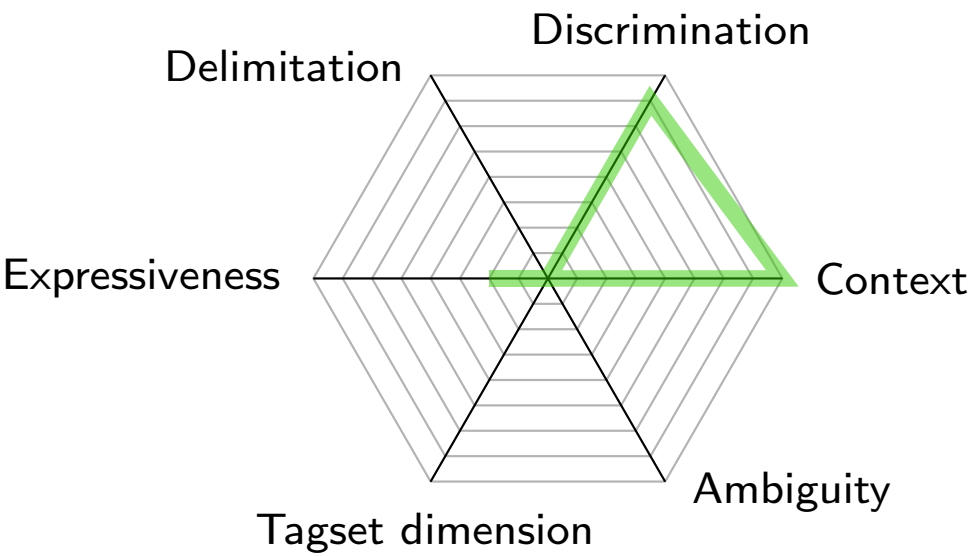
- number of **knowledge elements** to be rallied or degree of accessibility of the knowledge sources that are consulted:
  - annotation guidelines
  - nomenclatures (Swiss-Prot)
  - new sources to be found (Wikipedia, etc.)

Weight of the context

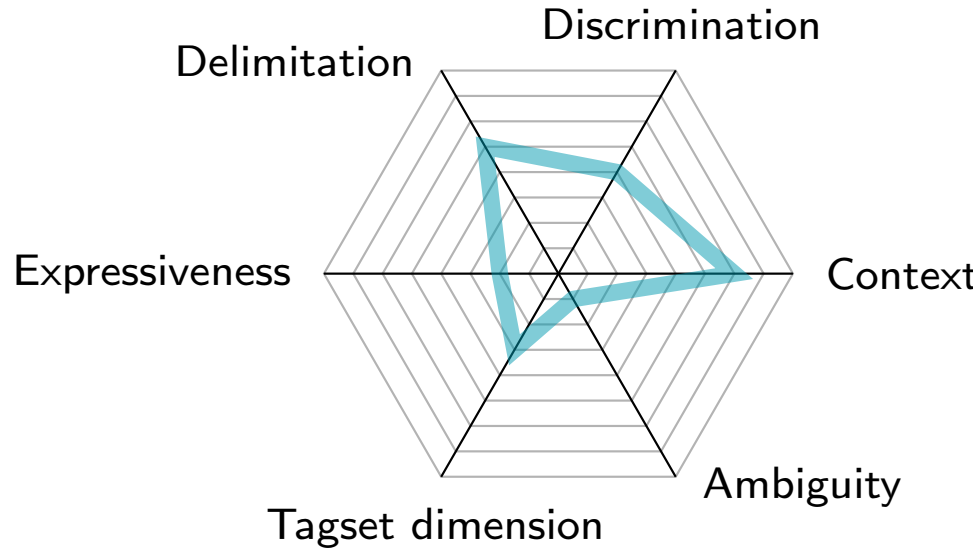


Where are the tools needed most?

Gene renaming relations



Structured named entities

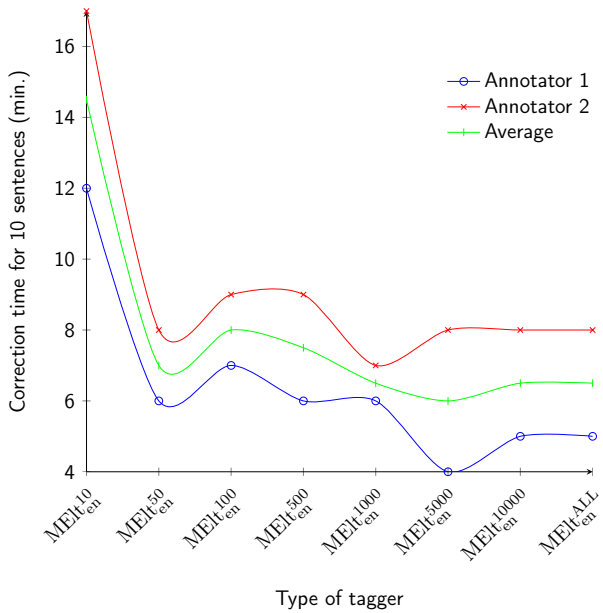


. . . according to the complexity profile of the campaign

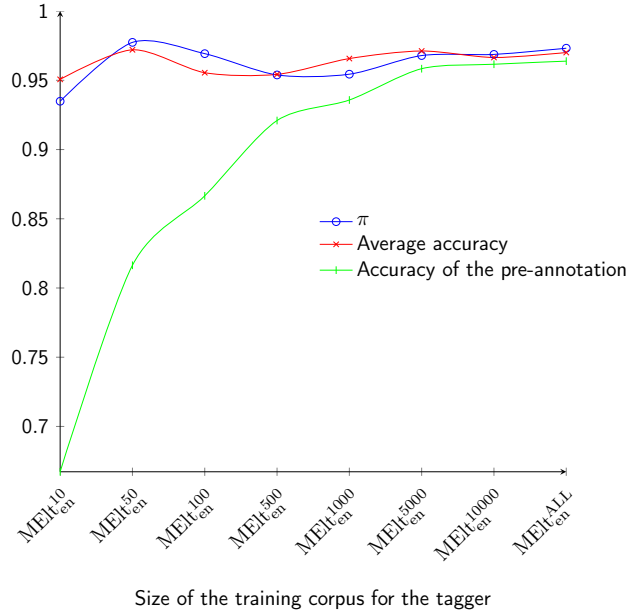
Impact of pre-annotation [Fort and Sagot, 2010]

- gain in time and in quality (inter-annotator agreement and accuracy)
- influence of the various levels of quality of the pre-annotation tool
- bias introduced by the pre-annotation

  ... while limiting the **effects of the learning curve**



(a) Correction time



(b) Correction quality

GWAPs

Promising:

- for language resources creation
- for a better understanding of the language resources creation process
  - decompose complexity
  - domain experts *vs* task experts (trained or not)
- **ethical**!

But:

- what can really be achieved is still unclear:
  - syntactic annotations? (we'll know soon!)
  - biomedical annotations?
- creating a "good" game still requires "talent" (ill-defined)
- quality evaluation remains an issue

Next steps: generalizing citizen science

- version 1.0 to come (mid-March)
- validation of the process and of the created resources
- other languages: English, German
- other, less-resourced, languages (Briton, Occitan, etc)

GWAPs platform for citizen science in France, driven by:

- ISC CNRS: Institut des Systèmes Complexes
- MNHN: Muséum national d'Histoire naturelle
- Paris-Sorbonne University


Abeillé, A., Clément, L., and Toussenel, F. (2003).
Building a treebank for French.
In Abeillé, A., editor, Treebanks, pages 165 –187. Kluwer, Dordrecht.


Artstein, R. and Poesio, M. (2008).
Inter-coder agreement for computational linguistics.
Computational Linguistics, 34(4):555–596.


Benzitoun, C., Fort, K., and Sagot, B. (2012).
TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe.
In Actes de Traitement Automatique des Langues Naturelles (TALN), pages 99–112, Grenoble, France.


Böhmová, A., Hajič, J., Hajičová, E., and Hladká, B. (2001).
The prague dependency treebank: Three-level annotation scenario.
In Abeillé, A., editor, Treebanks: Building and Using Syntactically Annotated Corpora. Kluwer Academic Publishers.

Bonneau-Maynard, H., Rosset, S., Ayache, C., Kuhn, A., and Mostefa, D. (2005).
Semantic annotation of the French Media dialog corpus.
In Proceedings of the InterSpeech, Lisbonne, Portugal.

Bontcheva, K., Cunningham, H., Roberts, I., and Tablan, V. (2010).
Web-based collaborative corpus annotation: Requirements and a framework implementation.
In Witte, R., Cunningham, H., Patrick, J., Beisswanger, E., Buyko, E., Hahn, U., Verspoor, K., and Coden, A. R., editors, Proceedings of the workshop on New Challenges for NLP Frameworks (NLPFrameworks 2010), La Valette, Malte. ELRA.

Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcade, M., and Poesio, M. (2013).
Using games to create language resources: Successes and limitations of the approach.

In Gurevych, I. and Kim, J., editors, The People's Web Meets NLP, Theory and Applications of Natural Language Processing, pages 3–44. Springer Berlin Heidelberg.

Chamberlain, J., Kruschwitz, U., and Poesio, M. (2009).
Constructing an anaphorically annotated corpus with non-experts: assessing the quality of collaborative annotations.
In Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources, People's Web '09, pages 57–62, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008).
Phrase Detectives: a web-based collaborative annotation game.
In Proceedings of the International Conference on Semantic Systems (I-Semantics'08), Graz, Autriche.

Cooper, S., Treuille, A., Barbero, J., Leaver-Fay, A., Tuite, K., Khatib, F., Snyder, A. C., Beenen, M., Salesin, D., Baker, D., and Popović, Z. (2010).

The challenge of designing scientific discovery games.
In Proceedings of the Fifth International Conference on the Foundations of Digital Games, FDG '10, pages 40–47, New York, NY, USA. ACM.


Fort, K., François, C., Galibert, O., and Ghribi, M. (2012a).
Analyzing the impact of prevalence on the evaluation of a manual annotation campaign.
In Proceedings of the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turquie.
7 pages.


Fort, K., Guillaume, B., and Chastant, H. (2014).
Creating Zombilingo, a Game With A Purpose for dependency syntax annotation.
In Gamification for Information Retrieval (GamifIR'14) Workshop, Amsterdam, Pays-Bas.


Fort, K., Nazarenko, A., and Rosset, S. (2012b).

Modeling the complexity of manual annotation tasks: a grid of analysis.
In Proceedings of the International Conference on Computational Linguistics (COLING), pages 895–910, Mumbaï, Inde.


Fort, K. and Sagot, B. (2010).
Influence of pre-annotation on POS-tagged corpus development.
In Proceedings of the Fourth ACL Linguistic Annotation Workshop, pages 56–63, Uppsala, Suède.


Geiger, D., Seedorf, S., Schulze, T., Nickerson, R. C., and Schader, M. (2011).
Managing the crowd: Towards a taxonomy of crowdsourcing processes.
In AMCIS 2011 Proceedings.


Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., and Quintard, L. (2011).
Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview.

In Proceedings of the 5th Linguistic Annotation Workshop, pages 92–100, Portland, Oregon, USA.
Poster.

Ide, N. and Romary, L. (2006).
Representing linguistic corpora and their annotations.
In Proceedings of the International Conference on Language Resources and Evaluation (LREC), Gène, Italie.

Khatib, F., DiMaio, F., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., Thompson, J., Popović, Z., et al. (2011).
Crystal structure of a monomeric retroviral protease solved by protein folding game players.
Nature structural & molecular biology, 18(10):1175–1177.

Kim, J.-D., Ohta, T., and Tsujii, J. (2008).
Corpus annotation for mining biomedical events from literature.
BMC Bioinformatics, 9(1):10.


Krippendorff, K. (2004).
Content Analysis: An Introduction to Its Methodology, second edition, chapter 11.
Sage, Thousand Oaks, CA., USA.


Lafourcade, M. and Joubert, A. (2008).
JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes.
In Actes de Journées internationales d'Analyse statistique des Données Textuelles (JADT), Lyon, France.


Lafourcade, M., Joubert, A., Schwab, D., and Zock, M. (2011).
Evaluation et consolidation d'un réseau lexical grâce à un assistant ludique pour le mot sur le bout de la langue.
In Actes de Traitement Automatique des Langues Naturelles (TALN), pages 295–306, Montpellier, France.


Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993).
Building a large annotated corpus of English : The Penn Treebank.

Computational Linguistics, 19(2):313–330.


Mathet, Y., Widlöcher, A., Fort, K., François, C., Galibert, O., Grouin, C., Kahn, J., Rosset, S., and Zweigenbaum, P. (2012).
Manual corpus annotation: Evaluating the evaluation metrics.
In Proceedings of the International Conference on Computational Linguistics (COLING), pages 809–818, Mumbaï, Inde.
Poster.


Pustejovsky, J. and Stubbs, A. (2012).
Natural Language Annotation for Machine Learning.
O'Reilly.


Stubbs, A. (2012).
Developing specifications for light annotation tasks in the biomedical domain.
In Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining, Istanbul, Turkey.


Tomanek, K., Hahn, U., Lohmann, S., and Ziegler, J. (2010).

A cognitive cost model of annotations based on eye-tracking data.
In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), ACL'10, pages 1158–1167, Stroudsburg, PA, USA. Association for Computational Linguistics.


Verspoor, K., Cohen, K. B., Lanfranchi, A., Warner, C., Johnson, H. L., Roeder, C., Choi, J. D., Funk, C., Malenkiy, Y., Eckert, M., Xue, N., Jr., W. A. B., Bada, M., Palmer, M., and Hunter, L. E. (2012).
A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools.
BMC Bioinformatics, 13:207.


Voormann, H. and Gut, U. (2008).
Agile corpus creation.
Corpus Linguistics and Linguistic Theory, 4(2):235–251.